Expert Systems with Applications 37 (2010) 3444–3453

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

BioDR: Semantic indexing networks for biomedical document retrieval

Anália Lourenço [a,*,1], Rafael Carreira [b,1], Daniel Glez-Peña [c,1], José R. Méndez [c,1], Sónia Carneiro [a], Luis M. Rocha [d,e], Fernando Díaz [f], Eugénio C. Ferreira [a], Isabel Rocha [a], Florentino Fdez-Riverola [c,1], Miguel Rocha [b,1]

[a] IBB – Institute for Biotechnology and Bioengineering, Centre of Biological Engineering, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal
[b] Department of Informatics/CCTC, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal
[c] Computer Science Department, University of Vigo, ESEI: Escuela Superior de Ingeniería Informática, Edificio Politécnico, Campus Universitario As Lagoas s/n, 32004 Ourense, Spain
[d] School of Informatics and Computing, Indiana University, 919 East Tenth Street, Bloomington, IN 47406, United States
[e] Instituto Gulbenkian de a Ciência, Apartado 14, 2781-901 Oeiras, Portugal
[f] Computer Science Department, University of Valladolid, Escuela Universitaria de Informática, Plaza Santa Eulalia, 9-11, 40005 Segovia, Spain

ARTICLE INFO

Keywords:
Biomedical document retrieval
Document relevance
Enhanced Instance Retrieval Network
Named Entity Recognition
Semantic indexing document network

ABSTRACT

In Biomedical research, retrieving documents that match an interesting query is a task performed quite frequently. Typically, the set of obtained results is extensive containing many non-interesting documents and consists in a flat list, i.e., not organized or indexed in any way. This work proposes BioDR, a novel approach that allows the semantic indexing of the results of a query, by identifying relevant terms in the documents. These terms emerge from a process of Named Entity Recognition that annotates occurrences of biological terms (e.g. genes or proteins) in abstracts or full-texts. The system is based on a learning process that builds an Enhanced Instance Retrieval Network (EIRN) from a set of manually classified documents, regarding their relevance to a given problem. The resulting EIRN implements the semantic indexing of documents and terms, allowing for enhanced navigation and visualization tools, as well as the assessment of relevance for new documents.

1. Introduction

Understanding the structure, dynamics, control and design of biological systems requires both theoretical and experimental information, mostly residing in scientific literature. Traditionally, researchers devoted a considerable amount of their time to the manual curation of literature, striving for the latest outcomes on a given subject. However, latest research has been prolific in publications and has led to an outstanding publishing rate.

Manual curation is now unfeasible. Database curators and biologists face a severe problem of information overload. Even when groups can afford to have people devoted to this task, they cannot keep up with collecting and analysing documents at the same pace as new research evolvements emerge. Furthermore, the ever evolving and often non-standardised biological terminology and the diverse and quite complex relationships among biological entities demand additional efforts from curators. The identification of important contents often implies searching for clues on unfamiliar and ambiguous terminology and thus, the ability to find very specific information in very large repositories and to cross-reference data adequately have become invaluable.

Available database contents should be taken into account, in particular, public access bibliographic engines and the ever growing number of Web-accessible open-access journals. Currently, PubMed is the bibliographic search system with the largest life science and biomedical coverage, used daily to perform thousands of queries over a repository with a few million documents. Query results are quite wealthy, since the repository provides document metadata such as title, authors and publishers, performs MeSH indexing, grants access to abstracts and sustains links to publisher's document location. However, PubMed does not sustain any result filtering, thus leading to a large set of undesired documents that increase the time and effort spent in further manual and/or automatic document processing.

Computational aid should embrace both automatic document retrieval and document relevance assessment, ensuring that the user is able to deploy a given query and obtain a reduced list of documents narrowed to the subject of analysis. Documents should

* Corresponding author. Address: IBB – Institute for Biotechnology and Bioengineering, Center of Biological Engineering, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal.
E-mail addresses: analia@deb.uminho.pt (A. Lourenço), rafaelcc@di.uminho.pt (R. Carreira), dgpena@uvigo.es (D. Glez-Peña), moncho.mendez@uvigo.es (J.R. Méndez), soniacarneiro@deb.uminho.pt (S. Carneiro), rocha@indiana.edu (L.M. Rocha), fdiaz@infor.uva.es (F. Díaz), ecferreira@deb.uminho.pt (E.C. Ferreira), irocha@deb.uminho.pt (I. Rocha), riverola@uvigo.es (F. Fdez-Riverola), mrocha@di.uminho.pt (M. Rocha).
[1] These authors contributed equally to this work.

0957-4174/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2009.10.044

be conveniently indexed, allowing intuitive document search, and far more important, sustaining focused searches based on the desired terminology. Thus, users would not only work over the subset of documents that they are actually interested in, but would also focus further reading and analysis based on mentions of genes, proteins and other biological entities that are meaningful within that context. In this regard, similar purpose approaches successfully applied to information retrieval and extraction in other domains should be looked into, assessing possible adaption.

In Fdez-Riverola, Iglesias, Díaz, Méndez, and Corchado (2007), we presented a successful spam filtering model called SPAMHUNTING, a Case-Based Reasoning (CBR) system which implements a disjoint knowledge representation engine able to address concept drift and disjoint category issues (Fdez-Riverola, Iglesias, Díaz, Méndez, & Corchado, 2007). In our previous experimentation using publicly available corpuses, we showed the superiority of the SPAMHUNTING system over well-known Machine Learning techniques, such as Support Vector Machines (SVM), different alternatives of the Naïve Bayes (NB) classifier, Boosting algorithms, Latent Semantic Indexing (LSI) and several approaches of lazy learning algorithms (Méndez, Corzo, Glez-Peña, Fdez-Riverola, Díaz, 2007; Méndez, Gonzalez, et al., 2007). Starting from our foregoing experience and following and analogy approach, we have adapted the methods and techniques used in the spam filtering domain to the present problem.

Therefore, the main contribution of this work is a novel semantic indexing approach, named BioDR, to enhance the retrieval of biomedical documents. Our final retrieval goal relates more directly to the needs of researchers using PubMed, i.e., we aim at delivering a tool that can assist end-users in their daily activities. As such, we addressed the filtering of PubMed's results, but we also provide for an indexing network that displays the documents according to user search perspectives, associating documents with similar contents and allowing term-specific views.

Throughout the next sections, we introduce PubMed as a major biomedical bibliographic search index and point out previous work on biomedical information retrieval. Next, using two real-world scenarios, we describe our indexing approach and explore the effect of different post-retrieval strategies on the filtering of documents returned by PubMed. In particular, we evaluate our approach over three sets of documents: raw abstracts retrieved from PubMed; the same abstracts, but after performing Named Entity Recognition (NER) for major biological entities; and a "best effort" dataset, where we perform NER over full-texts whenever available or over abstracts otherwise. For each scenario, we discuss document retrieval results and also show how the semantic indexing networks can be exploited by end-users. Finally, conclusions encompass final remarks and disclose future work.

2. Background

2.1. Indexing biomedical documents

Biomedical information retrieval is mostly supported by bibliographic databases and open-access journals. Currently, PubMed[2] sustains the largest biomedical bibliographic database, containing over 17 million records. It encompasses citations from biomedical literature as well as additional annotations (e.g. MeSH terms), abstract contents and full-text source links through LinkOut.[3] Moreover, it is fully linked with factual databases of DNA sequences, protein sequences and 3D molecular structures. The National Centre for Biotechnology Information (NCBI) provides access to its contents as part of the NCBI Entrez text-based search and retrieval system, allowing both online query and external access using the eUtils programming utilities.

Although providing an invaluable service, PubMed search engine is based on user-specified queries, i.e., sets of keywords that the user considers to best describe the problem. Achieving an adequate formulation of a query is not straightforward. Users may choose general terms or address broad-scope problems (e.g. a search on "leukemia" or "amino acid starvation"). While tracking down eventually relevant documents through such a process, many partially related and irrelevant documents will be retrieved as well.

Every document that matches the posted keywords in any of the requested search fields (e.g. title, keywords or abstract) is considered a candidate. However, it is not trivial for the user to pose a query in such a way that the keywords (single words or sets of words expected to co-occur together) do not bring attention over documents that are not connected to the subject of their interest.

For example, let us say that we are interested in searching for documents related to "*Escherichia coli* stringent response". If we impose the co-occurrence of the four words all together, we will most certainly miss many relevant documents due to discourse variants (e.g. "stringent response in *E. coli*" or "*E. coli*'s stringent response"). In turn, if we pose a word-free query, i.e., not imposing any word co-occurrence, we will get every document that matches any of our four query words. Probably, the wisest decision would be to re-structure the query, arranging the organism name "*E. coli*" and the event/problem "stringent response" as two search terms. Thus, it is ensured that the retrieved documents will refer to *E. coli* and be related to the desired subject. Yet, even then we may get a considerable number of partially related or irrelevant documents, for instance documents discussing stringent response in other organisms but alluding in any way to the microorganism *E. coli*.

2.2. Related work

Initiatives such as the KDD 2002 challenge cup,[4] BioCreAtIvE challenge[5] and TREC Genomics[6] have lead to related research. In KDD 2002, one of the tasks focused on helping to automate the work of curating biomedical databases by identifying what papers needed to be analysed for *Drosophila* gene expression information (Ghanem, Guo, Lodhi, & Zhang, 2002; Regev et al., 2002). The sub-task 2.3 of the BioCreAtIvE 2004 workshop addressed the automatic extraction and assignment of Gene Ontology (GO) annotations of human proteins, using full-text articles (Hirschman, Yeh, Blaschke, & Valencia, 2005). In turn, one task of the 2003 TREC Genomics Track was devoted to the collection of MEDLINE records for 50 gene topics (Hersh & Bhupatiraju, 2003) and in the 2004 TREC Genomics the same retrieval task embraced a broader variety of bioinformatics queries (Hersh et al., 2004).

Other works in the field address problems such as: the identification of protein interaction mentions using word proximity networks (Abi-Haidar et al., 2008; Verspoor et al., 2005); the ranking of MEDLINE abstracts based on the contents on the restriction enzyme database REBASE (Wilbur, 2000); the ranking of gene queries for the human genome (Sehgal & Srinivasan, 2006), the automatic classification of documents for the Immune Epitope Database (Wang, Morgan, Zhang, Sette, & Peters, 2007), the construction of content-rich biological networks (Chen & Sharp, 2004), the association of genes with Gene Ontology codes (Raychaudhuri, Chang, Sutphin, & Altman, 2002), the re-ranking of PubMed's results

[2] http://www.ncbi.nlm.nih.gov/pubmed/.
[3] http://www.ncbi.nlm.nih.gov/projects/linkout/.
[4] http://www.biostat.wisc.edu/~craven/kddcup/tasks.html.
[5] http://biocreative.sourceforge.net/.
[6] http://ir.ohsu.edu/genomics/.

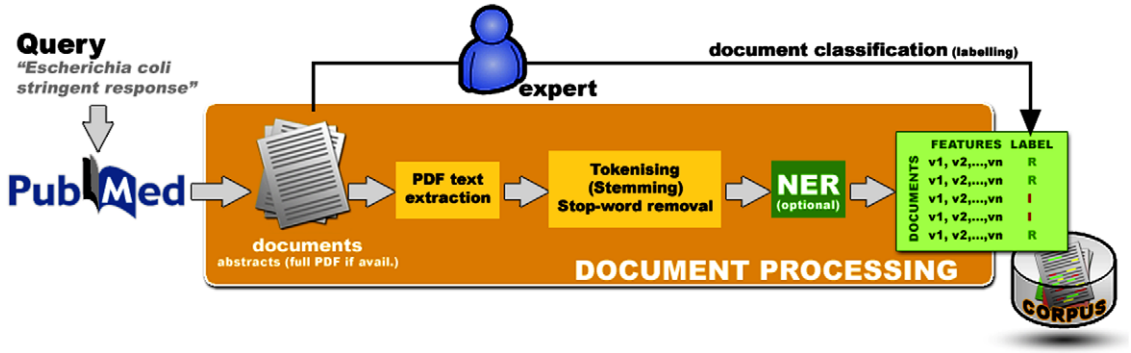

Fig. 1. Corpus construction at BioDR.

according to their relevance to SwissProt annotation (Dobrokhotov, Goutte, Veuthey, & Gaussier, 2003); and, the categorization of corpus of *Caenorhabditis elegans* papers (Mostafa & Lam, 2000).

Although most of these works have a very particular focus, it is interesting to notice that Machine Learning techniques (namely maximum entropy analysis, Naïve Bayes and probabilistic analysis) are currently combined with Natural Language Processing (NLP) techniques in order to tackle conventional linguistic analysis as well the particular biomedical terminology. Biomedical document retrieval is considered still an open research field, where PubMed results require further analysis in order to meet focused retrieval goals. We are also interested in improving retrieval performance.

Notwithstanding, our work differs from the literature above, in that we aim at delivering a rich document indexing network which, while focusing on relevant documents, provides means of navigation through the biological terms that best describe those documents. Users do not end up with a ranked list of documents, but rather a (semantic) network that can be traversed in an intuitive and useful way. The existence of different semantic layers, addressing major biological classes such as genes, proteins, compounds and organisms, allows users to switch analysis perspectives as appropriate.

3. Enhancing biomedical document retrieval with BioDR

3.1. Document retrieval and processing

Our workflow for document retrieval and processing encompassed three steps: retrieving documents from PubMed; pre-processing documents, namely performing PDF to text conversion and basic document structuring; and, applying a lexicon-based NER process (Fig. 1). Any tool that is able to perform such tasks and to output annotated documents can be used in this stage.

The only requirements are a NER module (lexicon-based or trained over gold standard corpora) and the tagging of major biological classes (namely, genes, proteins, compounds and organisms). In this work, the @Note Biomedical Text Mining open-source workbench (Lourenço et al., 2009),[7] a software platform developed by the authors, is used.

@Note supports PubMed search for relevant documents and document retrieval from open-access and subscribed Web-accessible journals. Entrez's eUtils grants access to PubMed and delivers the query results. Each PubMed record has a set of external links that the LWP (Library for WWW in *Perl*) crawling module follows to reach full-text documents (LWP, 2008). The original documents in PDF format are converted into plain ASCII files.

Plain text documents are tokenised and common English stopwords are filtered. Stemming was not considered here, because the NER module ensures term normalisation (i.e., all mention variants of a term are indexed by the common name). Our NER module is based on a dictionary obtained by merging the contents of some of the major biological databases (namely BioCYC, KEGG, EntrezGene and UniProt) and expert-specified lookup lists. A term rewriting system encompasses the set of active annotation rules, ranging from simple substitution rules to conditional and evaluated rules. Rules target up to seven-word terms and ignore too short words (less than 3 characters long). Furthermore, @Note sustains a user-friendly environment for the expert manual curation of document relevance.

3.2. Assessing the relevance of the terms

Taking as input the pre-processed and annotated set of documents generated in the previous phase, we are interested in selecting the most relevant terms representing entities belonging to the major biological classes (genes, proteins, compounds and organisms), for each document. Without any further information, the only way of doing this is to base it on the frequency of each word in the document. But, if we have available a collection of classified documents (a corpus), we can use information about the underlying distribution of the corpus in relation to the target concept (relevant or irrelevant) to assess the relevance of each term inside a specific document.

Therefore, we are interested in defining a criterion about the relevance of each term, T_j, which appears in a specific document, d, of a corpus K. In order to define this measure, the following reasoning is carried out. First, the probability that the document d is irrelevant (or belongs to class i) can be expressed as:

$$p(i|d) = \sum_{T_j \in d} p(i|T_j, d)p(T_j|d) \quad (1)$$

where $p(i|T_j, d)$ stands for the probability of a document d being irrelevant knowing that it contains the term T_j and $p(T_j|d)$ is the probability of occurrence of each term in the document.

The expression $p(T_j|d)$ is known, given the document d. Although the expression $p(i|T_j, d)$ is unknown, it can be estimated by the probability $p(i|T_j, K)$. That is, it can be approximated by the probability that a document in the corpus K is irrelevant if the term T_j is present in that document. Therefore, Exp. (1) can be approximated by:

$$p(i|d) \approx \sum_{T_j \in d} p(i|T_j, K)p(T_j|d) \quad (2)$$

Applying the Bayes' theorem, the probability $p(i|d)$ can be expressed as:

$$p(i|d) \approx \sum_{T_j \in d} \frac{p(T_j|i, K)p(i|K)}{p(T_j|K)} p(T_j|d) = p(i|K) \sum_{T_j \in d} \frac{p(T_j|i, K)p(T_j|d)}{p(T_j|K)} \quad (3)$$

[7] http://sysbio.di.uminho.pt/anote.php.

The probability that the document d is relevant (or belongs to class r) can be determined in a similar way:

$$p(r|d) \approx \sum_{T_j \in d} \frac{p(T_j|r,K)p(r|K)}{p(T_j|K)} p(T_j|r) = p(r|K) \sum_{T_j \in d} \frac{p(T_j|r,K)p(T_j|d)}{p(T_j|K)} \quad (4)$$

We are interested in discriminating between irrelevant and relevant terms (often, a term may have an approximately equal probability to appear in an irrelevant document as it does in a relevant one). Therefore, the relevance measure of a term should be able to identify highly predictive terms. This fact can be modelled by the difference between the Exps. (3) and (4), and each term of the sum can be interpreted as a measure of the contribution of each term in the final result, i.e., a measure of the relevance of each term. Moreover, if we are not interested in the sign of the contribution (positive if the term helps to classify a document as irrelevant or negative if it helps to classify one as relevant), the relevance of each term of the document can be defined as follows:

$$r(T_j,d) = \left\{ \frac{|p(i|K)p(T_j|i,K) - p(r|K)p(T_j|r,K)|}{p(T_j|K)} \right\} p(T_j|d) \quad (5)$$

The relevance measure $r(T_j,d)$ balances the local and global relevance of the term T_j. The first factor in $r(T_j,d)$ depends on the whole corpus K and expresses the utility of term T_j in order to discriminate among irrelevant or relevant documents and therefore it evaluates the global relevance of T_j. The second factor in $r(T_j,d)$ only depends on the specific document which is being processed and, hence, it can be viewed as a measure of the local relevance of T_j. As a consequence of this definition, the relevance of a term T_j which appears in two different documents only depends on the local relevance (since the first factor of Exp. (5) will be the same).

Moreover, the relative relevance of two terms T_j and T_k, which appear in a specific document d, not only depends on the local information, but also depends on the global information given by the first factor of Exp. (5), which will be probably different for both terms. This is particularly important because we are interested in ordering (by relative relevance) different terms in a specific document in order to select the most relevant ones.

Finally, this formulation can be used to select the most relevant terms in two ways: (i) a fixed number of terms ordered with respect to $r(T_j,d)$ or (ii) a variable number of terms depending on a fixed percentage of the whole sum of individual relevance values (if the terms of a document d are ordered descending by $|r(T_j,d)|$ and R is the sum of $|r(T_j,d)|$ over all the terms T_j belonging to d, then given a percentage α, the first k_α terms, whose partial sum of relevance values exceeds the quantity of αR, will be selected as the most relevant terms).

3.3. Indexing documents with the EIRN model

Based on the previous formulation for selecting relevant terms of each document in a corpus K, we present here our EIRN model for efficient and flexible document indexing and retrieval. The EIRN memory structure is borrowed from our previous successful SPAM-HUNTING system (Méndez, Glez-Peña, Fdez-Riverola, Díaz, & Corchado, 2008), from which we exploit its indexing capabilities and adaptive properties.

Following up the Case Retrieval Networks (CRN) indexing properties (Lenz, Auriol, & Manago, 1998), our model defines two measurements: (i) *term confidence* and (ii) *document confidence* for maintaining as much information as possible about existing data (terms and documents). Fig. 2 depicts an example of our EIRN model to document retrieval. The EIRN used in this work is characterized by a two-dimensional space, where the terms (cells) are organized according to the probability of representing irrelevant and relevant documents. Each cell in the network is associated with a *term confidence* (tc) which represents a measure of how much we can trust it to classify a given document. The value of tc for a given term T_j is given by Eq. (6).

$$tc_j = p(T_j|i,K) - p(T_j|r,K) \quad (6)$$

where $p(T_j|i,K)$ and $p(T_j|r,K)$ stand for the probability of the term T_j belonging to irrelevant and relevant documents, respectively.

The basic learning process in the EIRN consists in topology modification and term confidence adaptations. Based on a corpus K of training documents, learning in an EIRN is carried out by presenting all training documents to the network in a sequential fashion. For each training instance presentation, the network performs a so-called *learning cycle*, which may result in term confidence adaptation and topology modification.

In the first step of each learning cycle, the relevant terms (rt) of the actual input document d_m, are linked with the terms present in the network, adding new terms to the model if necessary. Each new connection is weighted up with a relevance value (rv_j) which represents the importance of this term to the actual document. The value of rv_j depends on the relevant terms (rt_m) of the input document d_m and the current term T_j. rv_j is calculated using:

$$rv_j = \frac{w_k}{2^{j-1}} \quad (7)$$

where w_k is a constant given by:

$$w_k = \frac{2^{rt_m - 1}}{2^{rt_m} - 1} \quad (8)$$

The second step consists in the adaptation of the term confidence affected in the previous step and the calculation of the actual *document confidence* (dc_m). The parameter dc represents a measure of document coherence by means of its relevant terms and aids in the identification of rare document contents. The value of dc for a given pair $\langle d_m, c_j \rangle$ is calculated by:

$$dc_m = \frac{\sum_{j=1}^{rt_m} p(T_j|c_j,K) rv_j}{rt_m} \quad (9)$$

where c_j represents the actual class of the document d_m, rt_m stands for the number of relevant terms for d_m, $p(T_j|c_j,K)$ represents the probability of the term T_j belonging to a document with the same class as document d_m and rv_j is calculated using Eq. (7).

3.4. Classifying new documents with BioDR

Every time a given document needs to be classified, the EIRN obtains a set M' composed of the documents most similar to the target document d'. In this sense, we can conceive the EIRN memory structure as a dynamic *k-nearest neighbour* mechanism able to retrieve a different number of neighbours depending on the terms selected from the unclassified document, d'. This is done by selecting the relevant terms of the new document as described in Section 3.2 and projecting them into the network term space (see Fig. 2). To perform this selection stage, the system encompasses two sequential steps: (i) calculating the distance between d' and the set of documents that share the greatest number of common terms (cf') and (ii) selecting those documents with a similarity value greater than the mean average value.

In order to calculate the similarity between two documents, given a set of shared relevant terms, we use a weighted distance metric that takes into account the relevance of each common term. The underlying idea is to weight those terms that are more

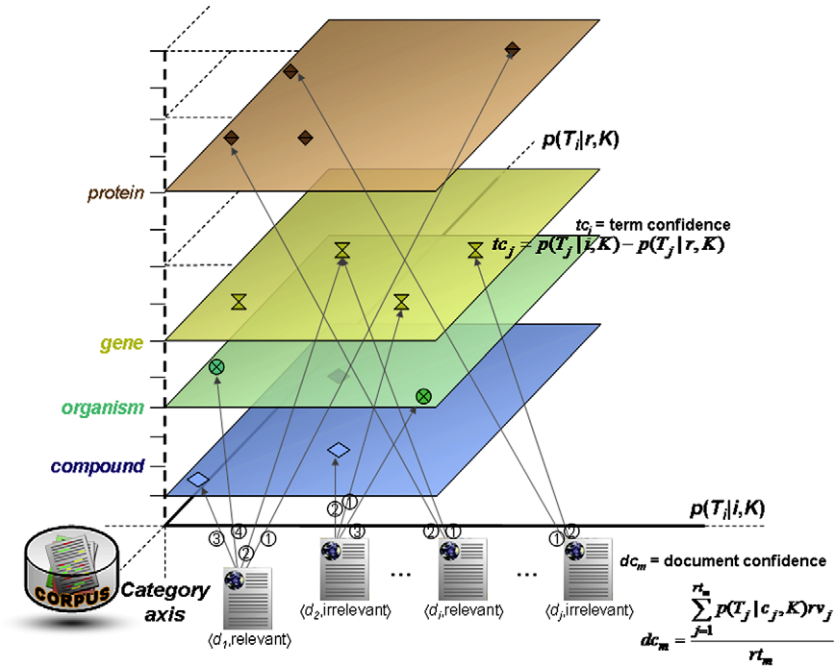

Fig. 2. EIRN architecture based on term annotation for document indexing and retrieval.

relevant to the target document d', using the position occupied by each of them in the arrows coming from the target document to the memory structure in Fig. 2. The value of the distance between the target document d' and a given document d_m is calculated by:

$$D(d', d_m) = \sum_{j=1}^{cf'} d(d'_j, d_{mj}) r v_j \quad (10)$$

where cf' is the number of common terms between M' and d', $r v_j$ represents the importance of each term to the target document d' and measures the distance between the position assigned to the common term j in the two documents, calculated as the difference between the situation of this term in the arrows coming from the target document d' and a given document d_m to the memory structure in Fig. 2.

Given the distance between two documents, the similarity is obtained by the following expression, where the document coherence is used to consider those texts which are most consistent with the corpus:

$$S(d', d_m) = \frac{1}{D(d', d_m)} dc_m \quad (11)$$

Every time BioDR executes the aforementioned document retrieval stage by selecting those documents with higher values for the similarity with the target document d', the system assigns a class label to the new document d' based on a proportional weighting voting algorithm. Each document in M' returns one vote and by means of recounting the existing votes, a final classification is provided by the system.

4. Model evaluation

4.1. Experimental setup

Our case studies concern research on the behaviour of the bacterium *E. coli* under particular stress conditions. In particular, Q1 represents the query *E. coli stringent response* and Q2 refers to the query *E. coli amino acid starvation*. Q1 keywords specifically address the *E. coli* stringent response while Q2 keywords aim at retrieving documents related to amino acid starvation, i.e., the condition that initiates the overall response to stress. Amino acid starvation triggers stringent response, while other conditions of starvation (e.g. nitrogen starvation) initiate other stress responses. Thus, any paper that addresses another starvation condition, but refers to amino acid starvation will be included in the results of Q2. Thus, Q1 is more particular than Q2 and this is reflected not only in the overall number of retrieved documents, but also on the number of irrelevant documents (Table 1). This situation allows us to test the performance of our model for imbalanced datasets and to discuss how it tackles the costs of erroneous classifications.

Table 1
Query statistics concerning document relevance, which was manually assessed by an expert curator.

	Relevant	Irrelevant	R:I ratio	Total
Q1	156 (55.1%)	127 (44.9%)	1.23	283
Q2	121 (34.4%)	231 (65.6%)	0.52	352

The performance of BioDR in document relevance classification is evaluated in two scenarios: considering abstracts with raw text (i.e. using all words in the text after pre-processing) and using only terms annotated in the NER process (i.e., only biological terms from the selected classes). Furthermore, the semantic indexing strategy is also evaluated in a third scenario, where NER is performed over full-texts, when these are available (abstracts are used in the other case).

4.2. Performance metrics

In a binary classification, the comparison between a predicted class and its actual class can be represented in a 2×2 contingency table (Table 2). Besides the accuracy of the classifier, given by $(TP + TN)/N$, where $N = TP + TN + FP + FN$, which represents the proportion of correctly classified documents, other commonly indicators of classifier performance can be derived from such 2×2 table

Table 2
The 2 × 2 contingency table. The abbreviations *TP*, *FP*, *FN*, and *TN* denote the number of respectively, true positives, false positives, false negatives, and true negatives.

		Actual class	
		Relevant (R)	Irrelevant (I)
Predicted class	Positive (+)	*TP*	*FP*
	Negative (−)	*FN*	*TN*

and they are defined either conditionally on actual class (recall or sensitivity, and specificity) or conditionally on predicted class (precision or positive predictive value, and negative predictive value).

The later indicators are defined as follows: (i) *recall* (also known as sensitivity or true positive rate, TPR) = $P(+|R) = TP/(TP + FN)$ which is the proportion of positive classifier results among the relevant documents, (ii) *specificity* = $P(-|I) = TN/(TN + FP)$ which represents the proportion of negative classifier results among the irrelevant documents (iii) *precision* (or *positive predictive value*, PPV) = $P\ (R|+) = TP/(TP + FP)$ which is the proportion of relevant documents among those with a positive predicted class, (iv) *negative predictive value* (NPV) = $P\ (I|-) = TN/(TN + FN)$ or proportion of irrelevant documents with a negative predicted class.

Unfortunately, none of these indicators validly represent the classifier discriminatory performance. Recall is only part of the discriminatory evidence (given the actual class), as high recall may be accompanied by low specificity. However, there are other interesting measures which try to give a global performance measure of the classifier, such as *f-score*, *kappa coefficient* and *diagnostic odds ratio* (DOR).

Firstly, *f-score* which was originally proposed by Rijsbergen (1979) to combine recall and precision (it ranges in the interval [0,1] and its value is 1 only if the number of FP and FN errors generated by the filter is 0):

$$f\text{-}score = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{12}$$

A variation of f-score is the so-called *balanced f-score* or $f\text{-}score_\beta$ (Shaw, Burgin, & Howell, 1997). As f-score, *balanced f-score* combines precision and recall but considering they have different importance. If $\beta = 1$ then precision and recall have the same weight, so f-score $= f\text{-}score_\beta$. If $\beta > 1$ then recall is more important than precision. Otherwise, precision has more weight. $f\text{-}score_\beta$ can be computed as Exp. (13) shows.

$$f\text{-}score_\beta = \frac{(\beta^2 + 1) \cdot precision \cdot recall}{\beta^2 \cdot precision + recall} \tag{13}$$

Secondly, the kappa coefficient, κ, is a more robust measure than the accuracy since it takes into account the agreement occurring by chance. This coefficient measures the agreement between two different classifiers, but in this work we assume that the actual class is given by a perfect classifier (which acts as an oracle), and then compare the target classifier with this one in order to measure its quality. The kappa coefficient is given by Exp. (14)

$$\kappa = \frac{Pr(o) - Pr(e)}{1 - Pr(e)} \tag{14}$$

where $Pr(o) = (TP + TN)/N$, is the observed agreement within the contingency table, and $Pr(e) = [(TN + FN)(TN + FP) + (FP + TP)(FN + TP)]/N^2$, is the expected agreement due to chance. The kappa coefficient has an upper-bound of 1, and a value equal to 1 indicates a total agreement.

Finally, the odds ratio used as single indicator of test performance is a third option. It does not depend on the *a priori* balance of probabilities among the relevant and irrelevant documents (the R:I ratio). The diagnostic odds ratio (DOR) of a classifier is the ratio of the odds of positivity in relevant documents relative to the odds of positivity in the irrelevant ones and it is given by Exp. (15).

$$\mathrm{DOR} = \frac{TP}{FN} \Big/ \frac{FP}{TN} \tag{15}$$

The value of DOR ranges from 0 to infinity, with higher values indicating better discriminatory classifier performance. If DOR is 1, it means that a classifier does not discriminate among relevant and irrelevant documents. Values lower than 1 point to improper classifier interpretation. The inverse of the DOR can be interpreted as the ratio of negativity odds within the relevant documents relative to the odds of negativity within the irrelevant ones. The DOR rises steeply when sensitivity or specificity becomes near perfect classification.

4.3. Evaluation of document retrieval performance

This section introduces our evaluation of the BioDR system using the performance metrics previously defined. For the experiments carried out in this work, we have used a 10-fold stratified cross-validation scheme (Kohavi, 1995), a technique that increases the confidence of experimental findings when using small datasets.

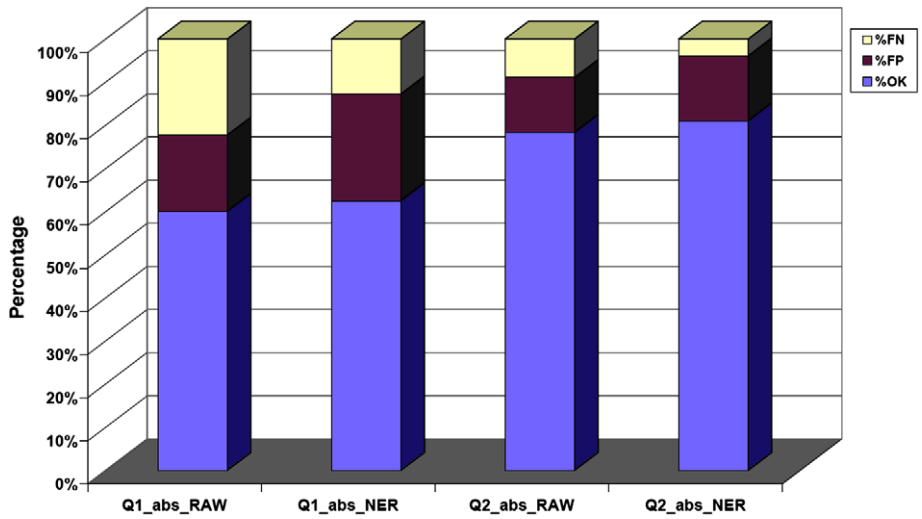

Fig. 3. Percentage of correct classifications, false positive and false negative errors for both queries using BioDR with 10-fold cross-validation.

With respect to the representation of each document, our EIRN was created using all the terms, capturing the maximum quantity of information ($\alpha = 100\%$).

Fig. 3 shows the percentage of correct classifications (*%TP + TN*), percentage of false positives (%FP) and percentage of false negatives (%FN) belonging to the two analysed queries. As we can see from Fig. 3, the proposed model drastically reduces the number of FN errors (relevant documents not detected) in both queries when the NER process is applied, when compared to the raw text-based analysis. Moreover, the system is able to achieve a higher accuracy.

Table 3 shows basic measures of the classifiers for both queries using BioDR, in the case of using raw text (*Q1_RAW*, *Q2_RAW*) and when performing NER (*Q1_NER* and *Q2_NER*). The first column shows the accuracy of the classifier and, as it can be observed, the classifier accuracy for the second query is better than the accuracy for the first one. For both queries, the use of NER improves slightly the accuracy and the recall, thus its use increases the proportion of well classified documents within the relevant documents. On the opposite, the proportion of well classified documents within the irrelevant documents (measured by the specificity) varies significantly for the first query and it is approximately the same for the second query. With regard to the predictive behaviour of the classifier, the use of NER barely changes the value of the precision of the classifier for both queries (it gets slightly worse for *Q1*, whereas it gets better for *Q2*). In the case of the negative predictive value, as it can be observed in Table 3, the use of NER improves in both queries its predictive value.

Table 3
Different performance results of the classifier: accuracy, recall (or sensitivity), specificity, precision (or positive predictive value) and negative predictive value for both queries using BioDR with 10-fold cross-validation.

	Accuracy	Recall (sensitivity)	Specificity	Precision (PPV)	NPV
Q1_RAW	0.60	0.60	0.61	0.65	0.55
Q1_NER	0.63	0.77	0.45	0.63	0.61
Q2_RAW	0.78	0.63	0.83	0.54	0.88
Q2_NER	0.81	0.84	0.80	0.57	0.94

However, the precision (and in the same way the negative predictive value) of the classifier for both queries is not comparable since this measure depends on the R:I ratio of the queries (and they are different for both queries). In order to show the effect of R:I ratio on the predictive values, Fig. 4a and c shows the extrapolated values of precision whereas Fig. 4b and d shows the estimated values of the negative prediction values for *Q1* and *Q2*, when the probability of relevant/irrelevant documents varies in the available corpus. Applying the Bayes's theorem, the precision of the classifier can be expressed as it is shown in Exp. (16).

$$precision(PPV) = P(R|+) = \frac{P(+|R)P(R)}{P(+|R)P(R) + P(+|I)P(I)} \tag{16}$$

For example, given Fig. 4a and c, and considering that for *Q1*, $P(R) = 0.55$ and for *Q2*, $P(R) = 0.24$, the estimated value of the precision for the inferred classifier from *Q1* (with the same recall and specificity) but with a $P(R) = 0.24$ (the observed probability of relevant documents in *Q2*), are 0.32 and 0.31 for the RAW and NER

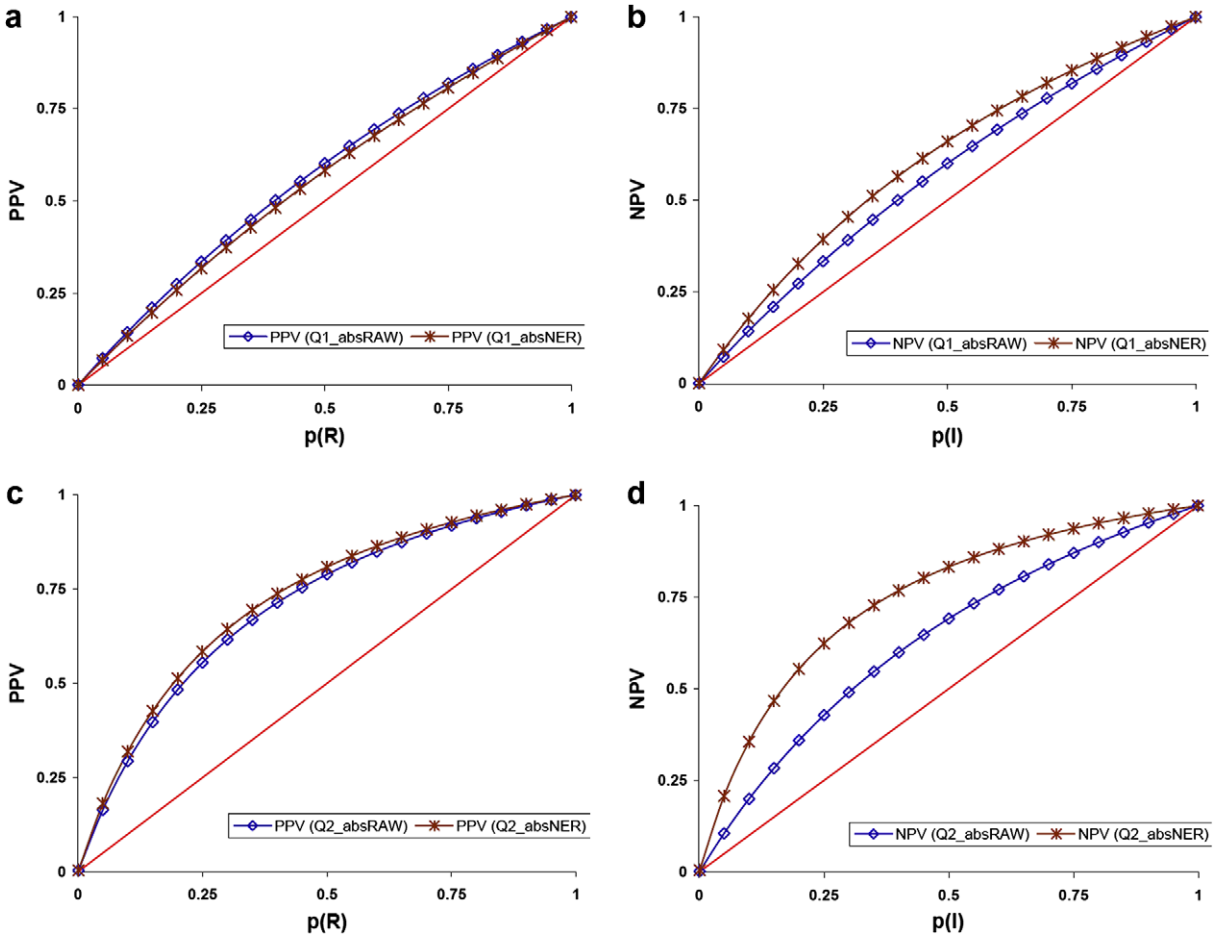

Fig. 4. BioDR model behaviour analysis for different scenarios of R:I query results.

Table 4
The f-score values for different balanced weights, kappa coefficient and diagnostic odds ratio for both queries using BioDR with 10-fold cross-validation.

	F-score			Kappa	DOR
	$\beta = 0.5$	$\beta = 1.0$	$\beta = 2.0$		
Q1_RAW	0.64	0.62	0.61	0.20	2.27
Q1_NER	0.66	0.69	0.74	0.22	2.71
Q2_RAW	0.56	0.58	0.61	0.44	8.38
Q2_NER	0.61	0.68	0.77	0.55	20.93

Table 5
Contribution of biological classes in the EIRN indexing structure.

Q1	Kappa	EIRN terms	EIRN terms	Kappa	Q2
(C)ompounds	0.31	8213	20,848	0.45	(C)ompounds
(P)roteins	0.18	2475	15,926	0.41	(G)enes
(O)rganisms	0.08	2707	14,290	0.38	(P)roteins
(G)enes	0.02	3723	13,321	0.02	(O)rganisms
(C + P)	0.24	10,688	36,774	0.49	(C + G)
(C + P + O)	0.23	13,395	51,064	0.51	(C + G + P)
(C + P + O + G)	0.22	17,118	64,385	0.55	(C + G + P + O)

classifier, respectively. Comparing these adjusted values for *Q1* with regard to the precision values of *Q2* it is not clear that the precision of the classifier for *Q1* was better than for *Q2*.

Consequently, and in order to avoid the effect of the R:I ratio and give a more robust performance measure of the classifier for both queries, Table 4 shows the $f\text{-}score_{\beta}$ (for three different weights of β), the kappa coefficient and the diagnostic odds ratio. The more meaningful measures from Table 4 are the kappa and DOR measures. In both queries these measures show that the use of NER improves the performance of the classifier, since kappa coefficient and DOR grows slightly for *Q1* and more significantly for *Q2*.

4.4. Biological assessment of the semantic indexing networks

In this section, we show the main capabilities of our EIRN indexing structure for the task of giving the final user access to retrieved documents. First of all, we discuss the importance of the biological classes of the terms identified by the NER process. Next, we describe the GUI that was implemented, showing the main features of the BioDR software.

In order to measure the contribution of each biological class in our EIRN indexing structure, Table 5 shows the individual value of the Cohen's Kappa coefficient for classification (using abstracts annotated with NER), as well as the total amount of terms stored in our EIRN model for each biological class (using full text where possible). As shown in Table 5, the biological class with a major impact in the model for both queries is "compounds" (higher value of the Kappa coefficient). In this sense, our model is able to correctly classify (using abstracts with NER) and efficiently index (using full

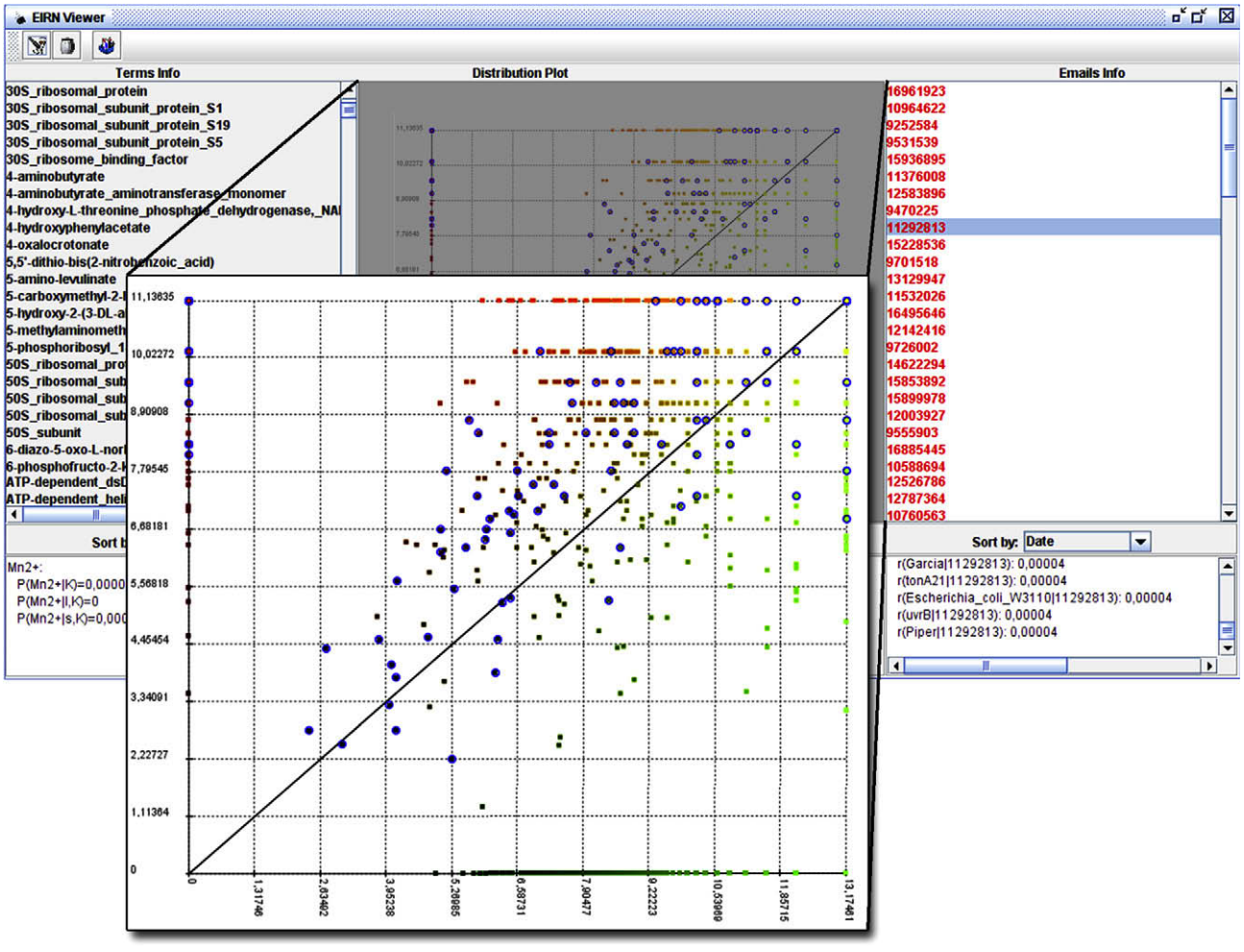

Fig. 5. The user interface of the EIRN.

text where possible) relevant documents with a percentage of terms below the 50% of the total amount.

Another interesting result from Table 5 is related with the direction of unbalanced queries. In the case of queries with a higher number of relevant documents (*Q*1), it is better to classify and index documents only with information coming from individual classes (i.e., compound or protein). However, in the situation in which irrelevant documents are more frequent (*Q*2), it is recommended to use as much information as possible.

Our EIRN can be analysed in a very intuitive way. Fig. 5 illustrates the network for the query *Q*2. Recurring to the EIRN viewer, we were able to identify the most predictive terms and analyse the biological associations with other terms. At left side, it lists the annotated terms and by clicking one term, the user visualises it in the plot. Likewise, by clicking the identifier of a processed document, presented at the right side, all dots representing its terms are marked (blue circles).[8] A colour gradient indicates the predictive ability of each term, ranging from red (relevant terms) to green (irrelevant terms). Terms that lay on the axis have the greatest predictive ability. *Y* axis stands for relevant terms while *X* axis represents irrelevant terms.

In advance, we knew that both queries should describe stress condition in a similar way, i.e., they describe the participants in the metabolic response to stress. Nevertheless, each stress condition has its characteristic terms, i.e., compounds, genes and enzymes that are particular to that event. The indexing network ranks terms by their predictive ability and links together related terms. Therefore, the user may analyse the most predictive terms of each query, assessing if the relationships established by the network are biologically meaningful.

For example, as expected, the most characteristic enzyme intervening on the stringent response of *E. coli*, '(p)ppGpp pyrophosphohydrolase', was found to be highly predictive as well as its enzymatic cofactor, Mn^{2+}. Genes like '*rel*A255' and '*rel*A256', known participants of the stringent response event, are also placed at the top of the *Y* axis. Similarly, terms that are known to be unrelated to *Q*2, like 'succinate_dehydrogenase', lay on the extreme side of the *X* axis.

5. Conclusions

The ability to find very specific information in very large repositories has become invaluable for the Biomedical research field. The retrieval of documents that match an interesting query is a task performed quite frequently and the manual revision of such results is laborious and time-consuming. The main contribution of this work is a novel approach to the enhanced retrieval of biomedical documents based on semantic indexing. The proposed approach supports the semantic indexing by identifying relevant terms in the documents based on a lexicon-based Named Entity Recognition process. It annotates occurrences of major biological classes (genes, proteins, compounds and organisms) in both abstracts and full-texts.

The system is based on a learning process that builds an EIRN from a set of manually classified documents, regarding their relevance to a given problem. The resulting EIRN implements the semantic indexing of documents and terms, allowing for enhanced navigation and visualization tools, as well as the assessment of relevance for new documents. Thus, users will not only work over the subset of documents that they are actually interested in, but also they will be able to focus further reading and analysis based on mentions to genes, proteins and other biological entities that are meaningful in a given context.

The proposed system was illustrated with two real-world queries related to research over *E. coli* stress response. The two queries present a different balance between relevant and irrelevant documents, thus imposing different challenges to the system. A number of performance measures were used to evaluate the system and to assess its behaviour when only NER terms are used, rather than the whole text. The results obtained in the task of classifying relevant documents are quite good, since the system is able to reduce significantly the number of irrelevant documents to be analysed, without a significant loss of relevant documents. Since the system provides probabilities, even lower values for false negatives could be obtained, if the user is willing to support a higher degree of false positives.

The analysis of the case studies has also shown the semantic indexing features of BioDR, as well as the tool that was developed to explore it. BioDR makes it possible to navigate semantically between documents and relevant terms, taking advantage of the rich contents of full-texts.

In future work, we aim at enhancing of the user interface towards the inclusion of new features for manual curation and visualization. Lexicon information about EIRNs terms and document mention hyperlinking are devised.

[8] For interpretation of colour in Fig. 5, the reader is referred to the web version of this article.

Acknowledgements

This work is partly funded by research project HUELLA (Ref. PS07/57) financed by the Consellería de Sanidade (Xunta de Galicia). The work of S.C. is supported by a Ph.D. Grant from the Fundação para a Ciência e Tecnologia (Ref. SFRH/BD/22863/2005). The work of D.G.P. is supported by a "Maria Barbeito" contract from Xunta de Galicia.

References

Abi-Haidar, A., Kaur, J., Maguitman, A., Radivojac, P., Retchsteiner, A., Verspoor, K., et al. (2008). Uncovering protein–protein interactions in the bibliome. *Genome Biology, 9*, S11.

Chen, H., & Sharp, B. M. (2004). Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics, 5* (October 8).

Dobrokhotov, P. B., Goutte, C., Veuthey, A. L., & Gaussier, E. (2003). Combining NLP and probabilistic categorisation for document and term selection for Swiss-Prot medical annotation. *Bioinformatics, 19*(90001), 91–94 (January).

Fdez-Riverola, F., Iglesias, E. L., Díaz, F., Méndez, J. R., & Corchado, J. M. (2007). SpamHunting: An instance-based reasoning system for spam labeling and filtering. *Decision Support systems, 3*(143), 722–736.

Fdez-Riverola, F., Iglesias, E. L., Díaz, F., Méndez, J. R., & Corchado, J. M. (2007). Applying lazy learning algorithms to tackle concept drift in spam filtering. *Expert Systems with Applications, 1*(33), 36–48.

Ghanem, M., Guo, Y., Lodhi, H., & Zhang, Y. (2002). Automatic scientific text classification using local patterns: KDD CUP 2002 (task 1). *ACM SIGKDD Explorations Newsletter, 4*(2), 95–96.

Hersh, W., & Bhupatiraju, R. T. (2003). TREC genomics track overview. In *Proceedings of the 12th text retrieval conference (TREC)* (pp. 14–23). Gaithersburg, MD:NIST. http://trec.nist.gov/pubs/trec12/papers/GENOMICS.OVERVIEW3.pdf.

Hersh, W., Bhupatiraju, R. T., Ross, L., Johnson, P., Cohen, A. M., & Kraemer, D. F. (2004). TREC 2004 genomics track overview. In *Proceedings of the 13th text retrieval conference (TREC)* (pp. 13–31).

Hirschman, L., Yeh, A., Blaschke, C., & Valencia, A. (2005). Overview of BioCreAtIvE: Critical assessment of information extraction for biology. *BMC Bioinformatics, 6*(Suppl. 1), S1.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th international joint conference on artificial intelligence* (Vol. 2, No. 12, pp. 1137–1143).

Lenz, M., Auriol, E., & Manago, M. (1998). Diagnosis and decision support. *Lecture Notes in Artificial Intelligence, 1400*, 51–90.

Lourenço, A., Carreira, R., Carneiro, S., Maia, P., Glez-Peña, D., Fdez-Riverola, F., et al. (2009). @Note: A Workbench for Biomedical Text Mining. *Journal of Biomedical Informatics, 42*(4), 710–720.

LWP::Simple – simple procedural interface to LWP (2008). Available from: URL: <http://search.cpan.org/~gaas/libwww-perl-5.810/lib/LWP/Simple.pm>.

Méndez, J. R., Corzo, B., Glez-Peña, D., Fdez-Riverola, F., & Díaz, F. (2007). Analyzing the performance of spam filtering methods when dimensionality of input vector

changes. In *Proceedings of the fifth international conference on machine learning and data mining*: *MLDM* (pp. 364–378).

Méndez, J. R., González, C., Glez-Peña, D., Fdez-Riverola, F., Díaz F, & Corchado, J. M. (2007). Assessing classification accuracy in the revision stage of a CBR spam filtering system. In *Proceedings of the seventh international conference on case-based reasoning:ICCBR* (pp. 374–388).

Méndez, J. R., Glez-Peña, D., Fdez-Riverola, F., Díaz, F., & Corchado, J. M. (2008). Managing irrelevant knowledge in CBR models for unsolicited e-mail classification. *Expert Systems with Applications*.

Mostafa, J., & Lam, W. (2000). Automatic classification using supervised learning in a medical document filtering application. *Information Processing and Management, 36*(3), 415–444.

Raychaudhuri, S., Chang, J. T., Sutphin, P. D., & Altman, R. B. (2002). Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. *Genome Research, 12*(1), 203–214.

Regev, Y., Finkelstein-Landau, M., Feldman, R., Gorodetsky, M., Zheng, X., Levy, S., et al. (2002). Rule-based extraction of experimental evidence in the biomedical domain: The KDD Cup 2002 (task 1). *ACM SIGKDD Explorations Newsletter, 4*(2), 90–92.

Rijsbergen, C. J. (1979). *Information Retrieval*.

Sehgal, A. K., & Srinivasan, P. (2006). Retrieval with gene queries. *BMC Bioinformatics, 7* (April 21).

Shaw, W. M., Burgin, R., & Howell, P. (1997). Performance standards and evaluations in IR test collections: Cluster-based retrieval models. *Information Processing and Management, 33*(1), 1–14.

Verspoor, K., Cohn, J., Joslyn, C., Mniszewski, S., Rechtsteiner, A., Rocha, L. M., et al. (2005). Protein annotation as term categorization in the gene ontology using word proximity networks. *BMC Bioinformatics, 6*(Suppl. 1), S20.

Wang, P., Morgan, A. A., Zhang, Q., Sette, A., & Peters, B. (2007). Automating document classification for the Immune Epitope Database. *BMC Bioinformatics, 8* (July 26).

Wilbur, W. J. (2000). Boosting Naive Bayesian learning on a large subset of MEDLINE. *Journal of the American Medical Informatics Association*, 918–922.